Procesamiento del Lenguaje Natural, Revista nº 61, septiembre de 2018, pp. 163-166 recibido 04-04-2018 revisado 26-04-2018 aceptado 04-05-2018

TUNER: Multifaceted Domain Adaptation for Advanced Textual Semantic Processing. First Results Available

TUNER: Adaptación a dominio para el procesamiento semántico avanzado. Primeros resultados disponibles

Rodrigo Agerri[1], Núria Bel[2], German Rigau[1], Horacio Saggion[2]
[1] University of the Basque Country (UPV/EHU)
[2] Pompeu Fabra University (UPF)
rodrigo.agerri@ehu.es; nuria.bel@upf.edu;
german.rigau@ehu.es; horacio.saggion@upf.edu

Abstract: The TUNER coordinated project (2016-2018) has focused on the development of domain adaptation technologies that reduce the cost of creating linguistic resources to develop systems in different languages and for different domains and genres. In this article we present the demonstrators, prototypes and resources that are already available project results.
Keywords: Language Resources, domain adaptation, semantic processing

Resumen: El proyecto coordinado TUNER (2016-2018) se ha centrado en el desarrollo de tecnologías de adaptación a dominio que permitan reducir el coste de la creación de recursos lingüísticos para desarrollar sistemas en diferentes lenguas y para diferentes dominios y géneros. En este artículo presentamos los demostradores, prototipos y recursos que son resultados del proyecto ya disponibles.
Palabras clave: Recursos lingüísticos, adaptación a dominio, procesamiento semántico

1 Introduction

TUNER project (http://ixa2.si.ehu.es/tuner/) started in 2016 as a coordinated project funded by the Spanish Ministry of Economy, Industry and Competitiveness. The project is coordinated by IXA Group[1] from University of the Basque Country (UPV/EHU) and the participant groups are: TALN Natural Language Research Group[2] and IULATERM-Technologies of Language Resources Group[3] from Pompeu Fabra University; TALG[4] Technologies and Applications of Galician Language, from University of Vigo; GRIAL[5], Linguistic Applications Inter-University Research Group, from University of Barcelona, and Open University of Catalonia; and Natural Language Processing & Information Retrieval Research Group[6] at National Distance Learning University (UNED). Researchers from Elhuyar Fundazioa[7] and Vicometch[8] are also participating in TUNER.

The project has focused on the research and development of domain adaptation technologies. These technologies were used for improving the Natural Language Processing (NLP) tools developed for the different languages in Spain and for different domains, in particular, health and tourism. These technologies have been applied to different tasks for which demonstrations have been built. In Section 3, some of these demonstrators are described, and in Section 4, the datasets

[1] http://ixa.eus/
[2] https://www.upf.edu/web/taln
[3] https://www.upf.edu/web/iulaterm/trl
[4] http://sli.uvigo.gal/
[5] http://grial.uab.es
[6] http://nlp.uned.es/web-nlp/
[7] https://www.elhuyar.eus
[8] http://www.vicomtech.org

ISSN 1135-5948. DOI 10.26342/2018-61-23

delivered by the project are described too to promote their reutilization.

2 TUNER objectives

In the general framework of text analytics and big data industry, the necessity of handling texts which are from different genres, domains and in different languages represents a very expensive investment in creating particular manually enriched datasets to train Natural Language Processing (NLP) methods. This dependency on data is challenging the expansion of companies to emerging business opportunities. TUNER addressed this problem through research and development of domain adaptation techniques to apply them to NLP technologies, to induce the information required to build different NLP tools for different tasks for which there are scarce or no data available.

TUNER research objectives were:

(i) Multilingual Text Processing. Providing language resources and tools for basic processing of textual data in the languages covered by the project: English, Spanish, Catalan, Basque and Galician.

(ii) Knowledge Acquisition and Integration. The improvement of a broad-coverage knowledge base integrated into the Multilingual Central Repository (MCR), linked to the latest versions of the English WordNet, and adapted to specific domains.

(iii) Reasoning and Inferencing. Development of inference and semantic reasoning engines using the knowledge integrated into the MCR.

(iv) Advanced Cross-Lingual Semantic Processing. Providing effective methodologies for developing robust and advanced semantic processing systems suitable to be adapted to other domains and languages.

3 Demonstrations and Prototypes

3.1 UMLS Mapper

This demonstrator, available at http://demos-v2.vicomtech.org/umlsmapper, shows the capabilities of a prototype (Perez, Cuadros and Rigau, 2018) to identify medical terms in free text in Spanish and map them to concepts of the UMLS medical terminology compendium[9]. The mapping is based on information retrieval techniques, by indexing these terminologies with Apache Lucene. Acronyms and abbreviations (Montoya, 2017) and IXA-pipes (Agerri, Bermudez and Rigau, 2014) were used to process the input texts. The disambiguation of terms with more than one possible mapping is based on UKB tool (Agirre and Soroa, 2009).

As for the demonstrator, at the main page you can choose between three given clinical texts or enter any free text to analyze. Once the analysis process is finished, two columns are displayed: on the right, the text sent is marked in several colors that highlight the recognized medical terms; on the left, the recognized terms are grouped by general medical concepts. By clicking on any of these terms, a third column shows additional information: hyponyms and hypernyms, synonymous expressions, a definition, etc.

3.2 AsisTerm

AsisTerm (http://scientmin.taln.upf.edu/scielo) is a prototype aimed to facilitate the understanding of complex biomedical terms in short texts --currently, abstracts of articles in English and Spanish from the ScieLO parallel corpus[10]-- by means of the identification and annotation of UMLS concepts and the automatic expansion of their definitions. Several resources, tools and methods have been developed to support and/or automate the semantic indexing of biomedical texts, but the vast majority of them are only targeted at English. When parallel corpora are available, as is the case of the ScieLO abstracts, these tools can be exploited in order to produce annotations in English that can then be transferred to texts in other languages. AsisTerm, in particular, makes use of the Becas online service[11] in order to automatically annotate the ScieLO English abstracts with UMLS concepts. Once the relevant concepts are identified, the spans of texts in the Spanish abstracts that best match each of them have to be determined. This is done by computing the similarity between candidate terms obtained from the abstract in Spanish and the Spanish lexicalizations of each retrieved UMLS concept. Once the correct location for each annotation is found, the definition of the corresponding concept is expanded by means of the MedlinePlus Connect service[12]. The system includes a web

[9] https://www.nlm.nih.gov/research/umls

[10] http://www.scielo.org

[11] http://bioinformatics.ua.pt/becas/

[12] https://medlineplus.gov/connect

interface to search and retrieve the ScieLO abstracts in English and Spanish where the annotations and definitions of biomedical terms can be visualized.

3.3 Lingaliza

Lingaliza (http://sli.uvigo.gal/lingaliza/) is a web interface designed to test the full new set of NLP tools provided by the IXA pipes for the Galician language. This set of tools includes a rule-based tokenizer and sentence segmenter, a statistical lemmatizer and POS tagger, a state-of-the-art NER tagger a wikification tool based on DBpedia Spotlight, a NED tool based on DBpedia Spotlight and a UKB graph-based tool for word sense disambiguation.

3.4 Analhitza

Analhitza is a web interface for easy access to the NLP tools provided by the ixaKat for Basque and IXA pipes for Spanish languages, among others (Otegi et al., 2017). Analthitza (http://ixa2.si.ehu.es/clarink/analhitza.php) uses the basic IXA NLP tools, including segmentation, lemmatization and POS and NER tagging.

3.5 Central Unit Detector for Scientific Texts

With the aim of building an automatic discourse analyzer, we developed a tool consisting of a discourse segmenter and an automatic central discourse unit detector. In the future, we will link the segments with discourse relations, on the top of the central unit (CU). The CU detector, for Basque (Bengoetxea, Atutxa and Iruskieta, 2017) (http://ixa2.si.ehu.es/CU-detector/) and for Spanish (Bengoetxea and Iruskieta, 2018) detects the most salient node of a text (http://ixa2.si.ehu.es/clarink/tools/ES-CU-detector/). First, the detector segments the text in discourse units and then it labels the most important text segment.

The CU detector for Basque was developed to handle scientific abstract texts from seven different specialized domains: medicine, health. life, terminology, science, economy and computer science. Scientific texts from psychology and linguistics domains were employed to develop the Spanish CU detector.

3.6 Navigating the SEPLN Anthology

As part of our work on scientific text mining, we have developed a tool to process the articles from the SEPLN journal effectively creating the *SEPLN Anthology*, a fully analyzed bi-lingual resource created from SEPLN publications (Saggion et al., 2017) available at http://scipub-taln.upf.edu/sepln/. Furthermore, we have also developed a Web-based information access platform which exploits the SEPLN Anthology documents to provide single-document and collection-based visualizations as a means to explore the rich generated contents.

3.7 PDFdigest: Extracting content from PDF

We have developed PDFdigest (Ferrés et al., 2018), a tool for PDF-to-XML content extraction specially designed to extract scientific articles' headings and logical structure (title, authors, abstract, etc.) and its textual content (http://taln.upf.edu/pdfdigest). PDFdigest has been used to extract metadata from the SEPLN journal to create the SEPLN Anthology (Saggion et al., 2017).

4 Resources

4.1 Galnet

Galnet is an open wordnet for Galician, aligned with the interlingual index (ILI) generated from the English WordNet3.0, following the expand model for the creation of new wordnets, where the variants associated with the Princeton WordNet synsets are translated using different strategies. Galnet can be searched via a web interface (http://sli.uvigo.gal/galnet/) and can be downloaded in RDF and LMF formats from http://sli.uvigo.gal/download/SLI_Galnet/.

Galnet is part of the Multilingual Central Repository (http://adimen.si.ehu.es/web/MCR/) a database that currently integrates wordnets from six different languages (English, Spanish, Catalan, Galician, Basque and Portuguese) using WordNet 3.0 as ILI and where each *synset* is classified under the WordNet Domains hierarchy, the SUMO ontology and the Top Concept Ontology. The specific interface designed to query Galnet extends the MCR functionalities by providing different navigation types through domain hierarchies and ontologies, allowing an interactive tree-based visualization of *synsets* by their semantic relations, including a variety of terminology-oriented functionalities.

4.2 SensoGal Corpus

The SensoGal Corpus is an English-Galician parallel corpus semantically annotated with respect to WordNet 3.0 and aligned at the sentence and word level with the English SemCor corpus. The SensoGal Corpus is based on the Galician translation of the 186 fully tagged texts included in the original English Princeton SemCor, prioritizing the translation of the texts available in the MultiSemCor Corpus. The resulting parallel corpus can already be consulted through a web query interface http://sli.uvigo.gal/SensoGal/. The Galician corpus sentences in SensoGal are used as lexicographic examples of usage of the variants in the Galnet interface.

4.3 IULA Spanish Clinical Record Corpus

The IULA Spanish Clinical Record Corpus (IULA-SCRC) is a corpus of 3,194 sentences extracted from anonymized clinical records and manually annotated with negation markers and their scope. Annotation guidelines are documented at Marimon, Vivaldi and Bel (2017) and the corpus access is http://eines.iula.upf.edu/brat/#/NegationOnCR_IULA.

Acknowledgments

TUNER coordinated project was funded by Spanish Ministry of Economy, Industry and Competitiveness (TIN2015-65308-C5, MINECO/FEDER, UE).

References

Agerri, R., J. Bermudez, and G. Rigau (2014). IXA pipeline: Efficient and Ready to Use Multilingual NLP tools. In Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014), pp. 3823-3828.

Agerri, R.; X. Gómez Guinovart; G. Rigau; M. A. Solla Portela (2018). Developing New Linguistic Resources and Tools for the Galician Language. Proceedings of the 11th Language Resources and Evaluation Conference (LREC'18).

Agirre, E. and A. Soroa. (2009). Personalizing PageRank for Word Sense Disambiguation. In Proceedings of the 12th Conference of the European Chapter of the ACL, pp. 33-41.

Bengoetxea, K.; A. Atutxa; M. Iruskieta (2017). Un detector de la unidad central de un texto basado en técnicas de aprendizaje automático en textos científicos para el euskera. Procesamiento del Lenguaje Natural, 58, 37-44.

Bengoetxea, K.; M. Iruskieta (2018). A Supervised Central Unit Detector for Spanish. Procesamiento del Lenguaje Natural, 60, 29-36.

Ferrès, D.; Saggion, H.; Rozano, F.; Bravo, À. (2018). PDFdigest: an Adaptable Layout-Aware PDF-to-XML Textual Content Extractor for Scientific Articles. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).

Gómez Guinovart, X.; M. A. Solla Portela (2018). Building the Galician wordnet: methods and applications. Language Resources and Evaluation, 52:1, 317-339.

Marimon, M., J. Vivaldi, and N. Bel. (2017). Annotation of negation in the IULA Spanish clinical record corpus. In Proceedings of the Workshop SemBEaR 2017. ACL. p. 43-52.

Montoya, I. (2017). Análisis, normalización, enriquecimiento y codificación de historia clínica electrónica (HCE). Tesis del Máster Universitario Konputazio Ingeniaritza eta Sistema Adimentsuak, (UPV/EHU).

Otegi, A.; O. Imaz; A. Díaz de Ilarraza; M. Iruskieta; L. Uria (2017). ANALHITZA: a tool to extract linguistic information from large corpora in Humanities research. Procesamiento del Lenguaje Natural 58: 77-84.

Perez, N., M. Cuadros y G. Rigau (2018). Biomedical term normalization of EHRs with UMLS. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).

Saggion, H.; F. Ronzano; P. Accuosto; D. Ferrés (2017). MultiScien: a Bi-Lingual Natural Language Processing System for Mining and Enrichment of Scientific Collections. BIRNDL@SIGIR (1) 2017: 26-40

Solla Portela, M. A. and X. Gómez Guinovart (2017). Diseño y elaboración del corpus SemCor del gallego anotado semánticamente con WordNet 3.0. Procesamiento del Lenguaje Natural, 59, 137-140 (ISSN 1135-5948